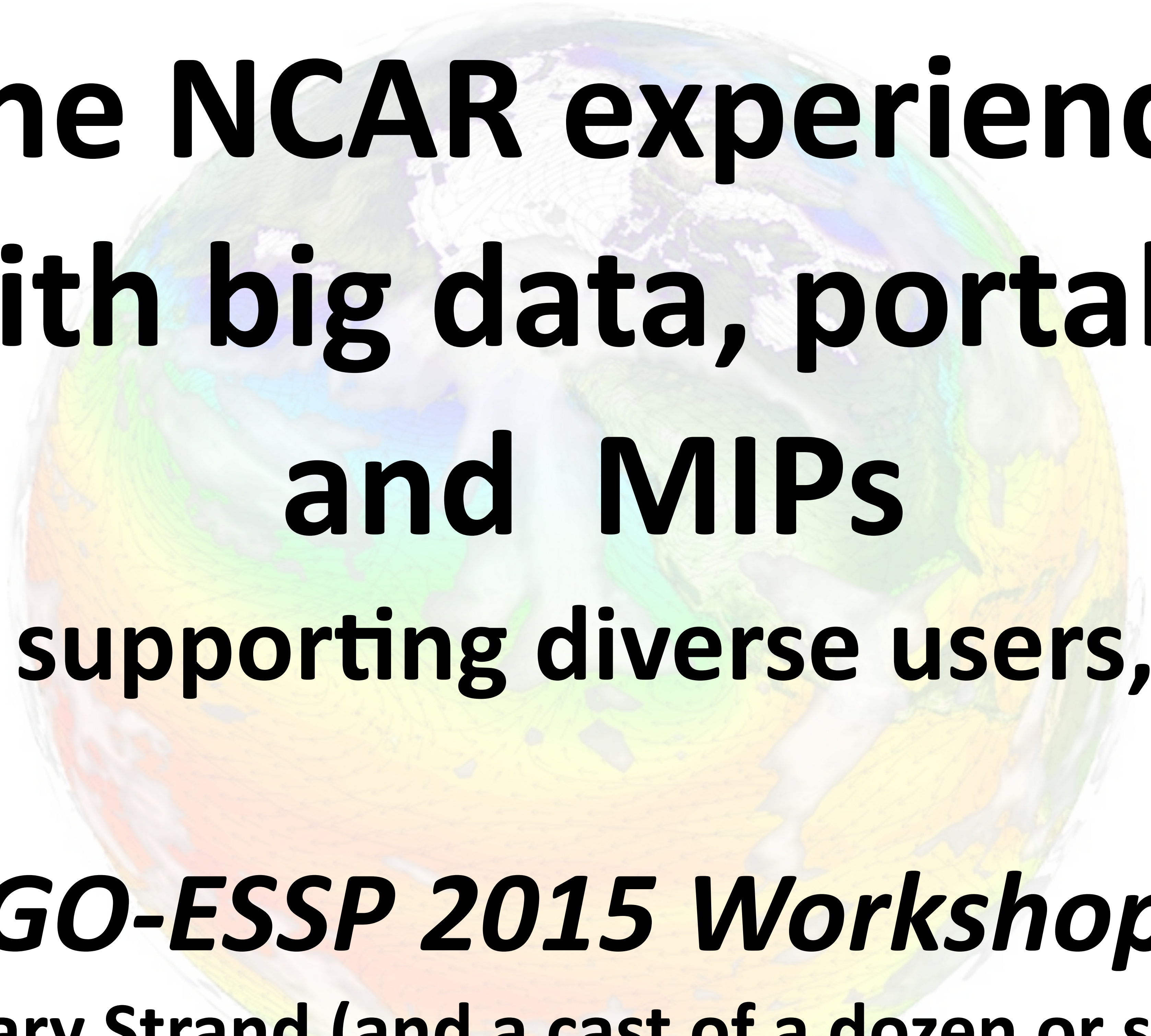


The NCAR experience with big data, portals, and MIPs

(and supporting diverse users, too)

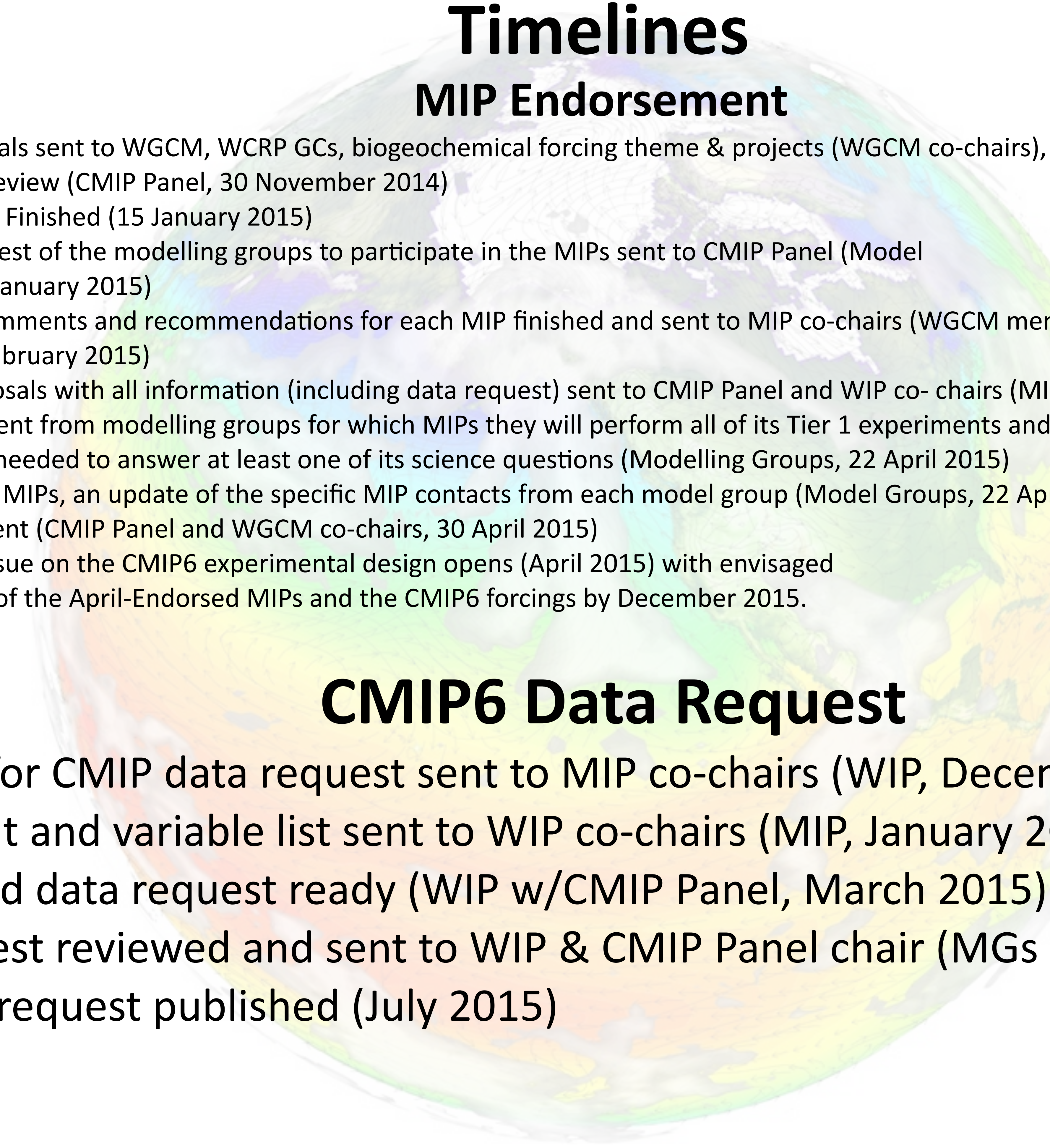
GO-ESSP 2015 Workshop

Gary Strand (and a cast of a dozen or so)

strandwg@ucar.edu

NCAR CMIP3 and CMIP5 comparison

Category	CMIP3	CMIP5
Models used	**2:** CCSM3 & PCM	**5:** CCSM4, CESM1-CAM5, CESM1-BGC, CESM1-WACCM, CESM1-FASTCHEM
Total volume submitted	**~ 9.2 TB**	**~175 TB**
Total volume generated	**~120 TB**	**~1,400 TB**
Total simulated years	**~14,900**	**~28,500**
Number of model runs	107 total 73 (CCSM3) 34 (PCM1)	555 total 91 (CCSM4 long-term) 400 (CCSM4 DP 64 (other configurations)
Experiments requested	12	37
Output categories	6	19
Number of requested fields	137	951
Persons actively involved	10	15
Months start-finish	36 (2004-2006)	48 (2010-2013)



Timelines

MIP Endorsement

- Revised proposals sent to WGCM, WCRP GCs, biogeochemical forcing theme & projects (WGCM co-chairs), MIP co-chairs and modelling groups for review (CMIP Panel, 30 November 2014)
- Review Process Finished (15 January 2015)
- Update of interest of the modelling groups to participate in the MIPs sent to CMIP Panel (Model Groups, 15 January 2015)
- Synthesis of comments and recommendations for each MIP finished and sent to MIP co-chairs (WGCM members organized by WGCM co-chairs, 15 February 2015)
- Final MIP proposals with all information (including data request) sent to CMIP Panel and WIP co- chairs (MIP co-chairs, 31 March 2015)
- Firm Commitment from modelling groups for which MIPs they will perform all of its Tier 1 experiments and providing all the requested diagnostics needed to answer at least one of its science questions (Modelling Groups, 22 April 2015)
- For each of the MIPs, an update of the specific MIP contacts from each model group (Model Groups, 22 April 2015)
- MIP Endorsement (CMIP Panel and WGCM co-chairs, 30 April 2015)
- GMD Special Issue on the CMIP6 experimental design opens (April 2015) with envisaged submission of the April-Endorsed MIPs and the CMIP6 forcings by December 2015.

CMIP6 Data Request

Template for CMIP data request sent to MIP co-chairs (WIP, December 2014)

Experiment and variable list sent to WIP co-chairs (MIP, January 2015)

Synthesized data request ready (WIP w/CMIP Panel, March 2015)

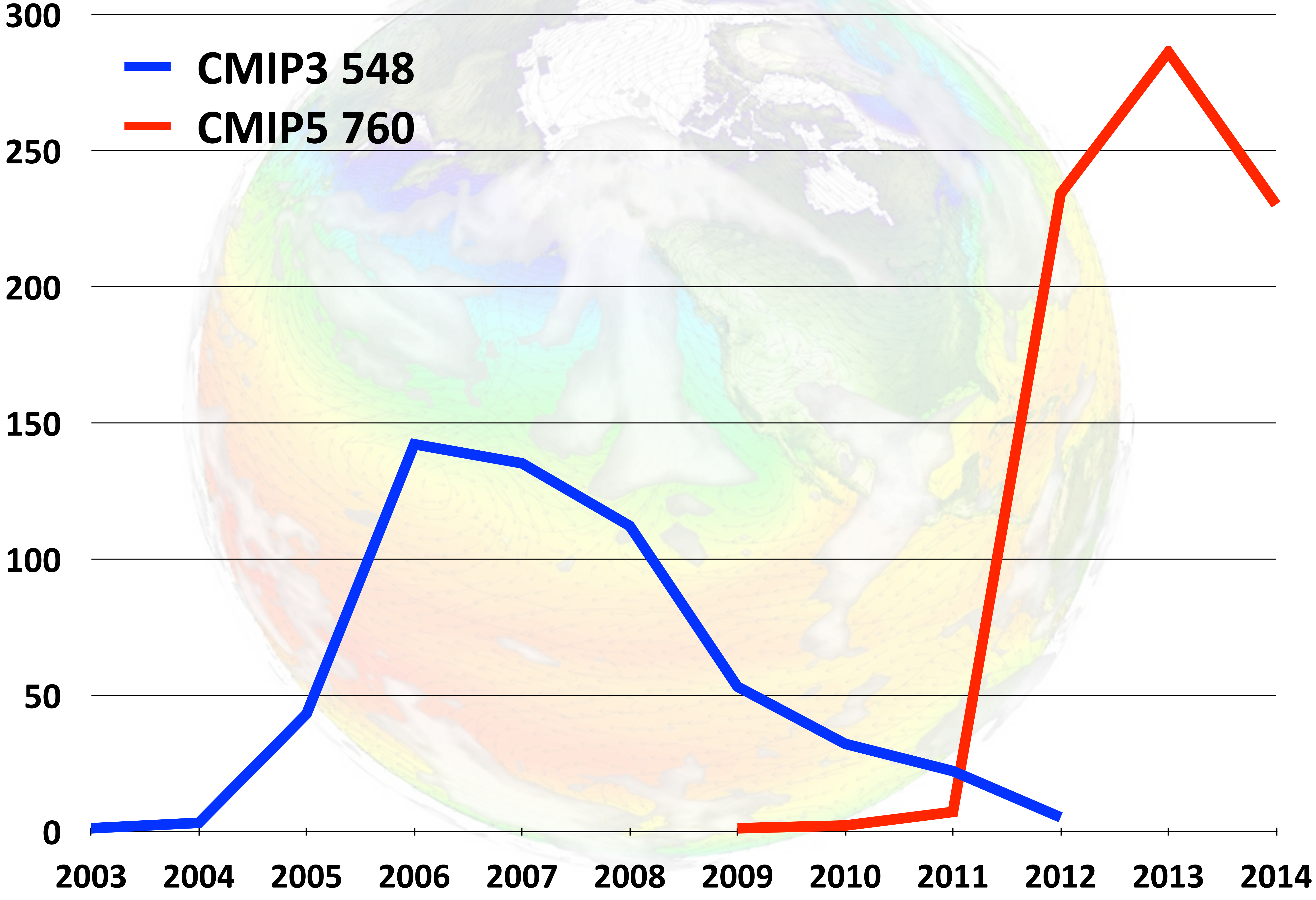
Data request reviewed and sent to WIP & CMIP Panel chair (MGs & MIP, April 2015)

Final data request published (July 2015)

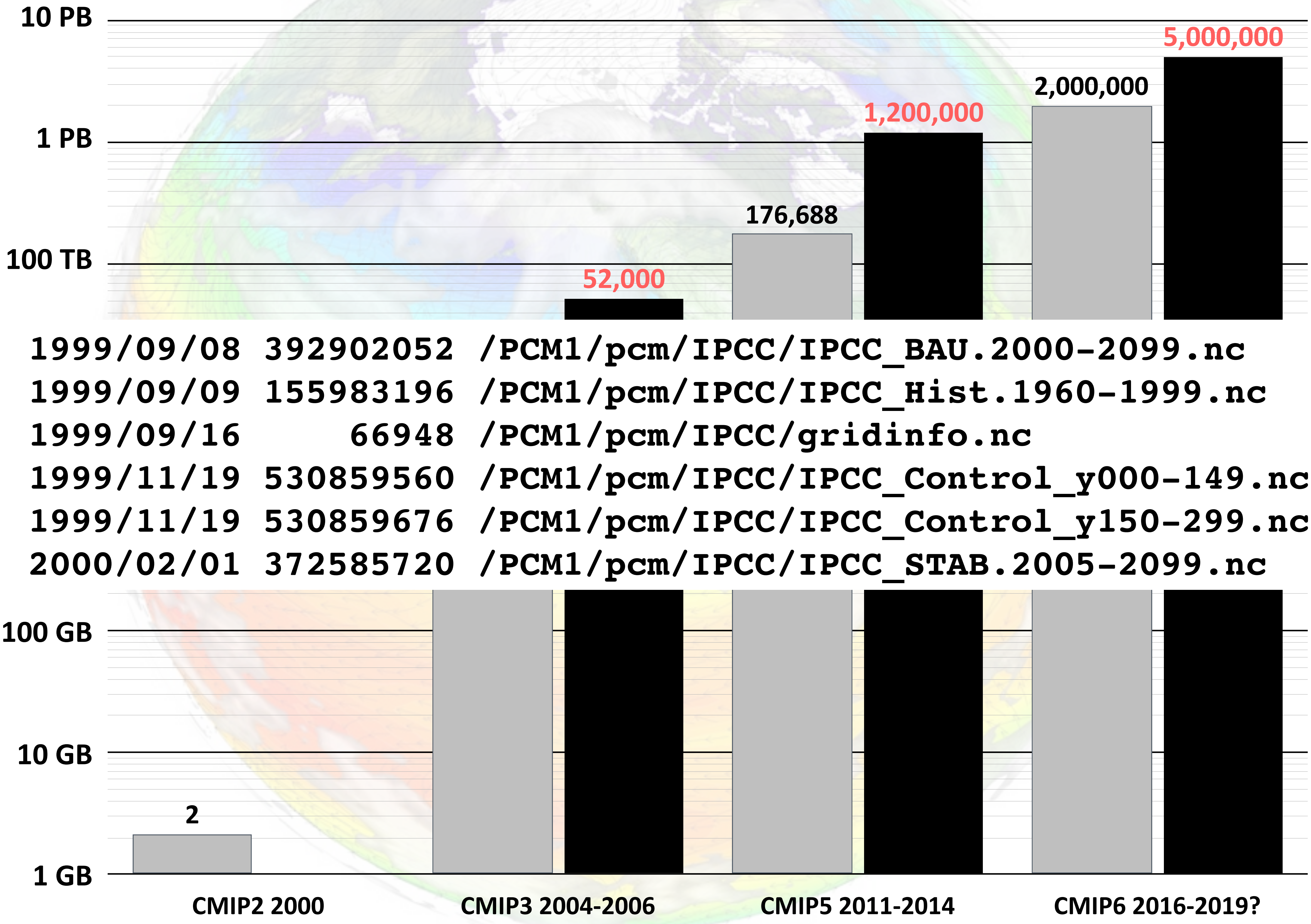
Overview of MIPs that have applied for CMIP6 Endorsement

Short Name	Long Name	Short Name	Long Name
AerChemMIP	Aerosols and Chemistry MIP	**LS3MIP**	Land Surface, Snow and Soil Moisture
C4MIP	Coupled Climate Carbon Cycle MIP	**LUMIP**	Land Use MIP
CFMIP	Cloud Feedback MIP	**OCMIP6**	Ocean Carbon Cycle MIP, Phase6
DAMIP	Detection and Attribution MIP	**OMIP**	Ocean MIP
DCPP	Decadal Climate Prediction Project	**PDRMIP**	Precipitation Driver and Response MIP
ENSOMIP	ENSO MIP	**PMIP**	Paleoclimate
FAFMIP	Flux Anomaly Forced MIP	**RFMIP**	Radiative Forcing MIP
GeoMIP	Geoengineering MIP	**ScenarioMIP**	Scenario MIP
GMMIP	Global Monsoons MIP	**SolarMIP**	Solar MIP
HighResMIP	High Resolution MIP	**VolMIP**	Volcanic Forcings MIP
ISMIP6	Ice Sheet MIP		
DiagnosticMIPs **(no proposed experiments rather requesting that certain output is archived and/or contributing to the evaluation and analysis in a coordinated manner)**			
CORDEX	Coordinated Regional Climate Downscaling Experiment		
DynVar	Dynamics and Variability of the Stratosphere Troposphere System		
GDDEX	Global Dynamical Downscaling Experiment		
SIMIP	Sea Ice MIP		
VIAAB	VIA AdvisoryBoardforCMIP6		

Publications by year

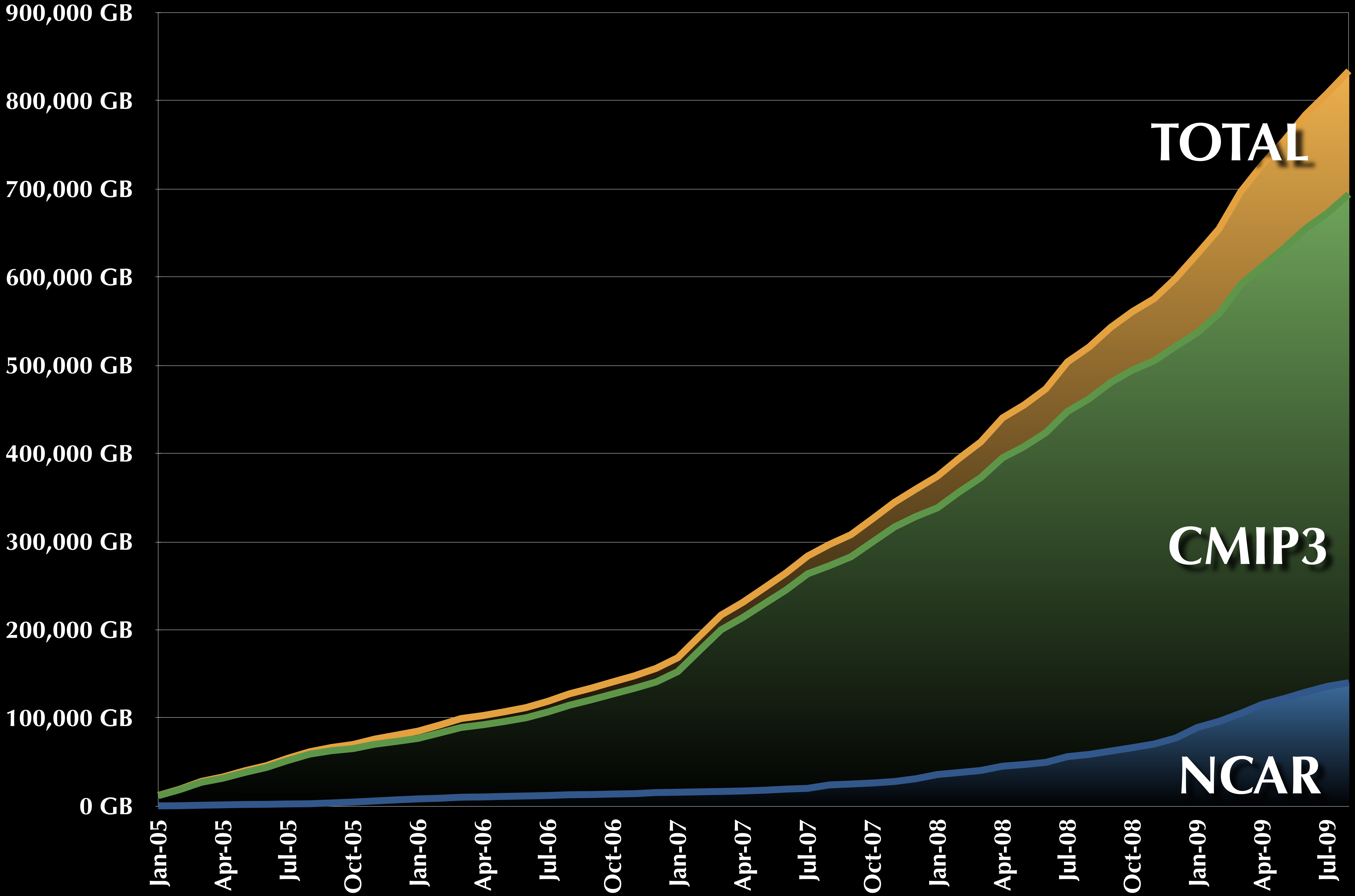


NCAR contributions to CMIPs

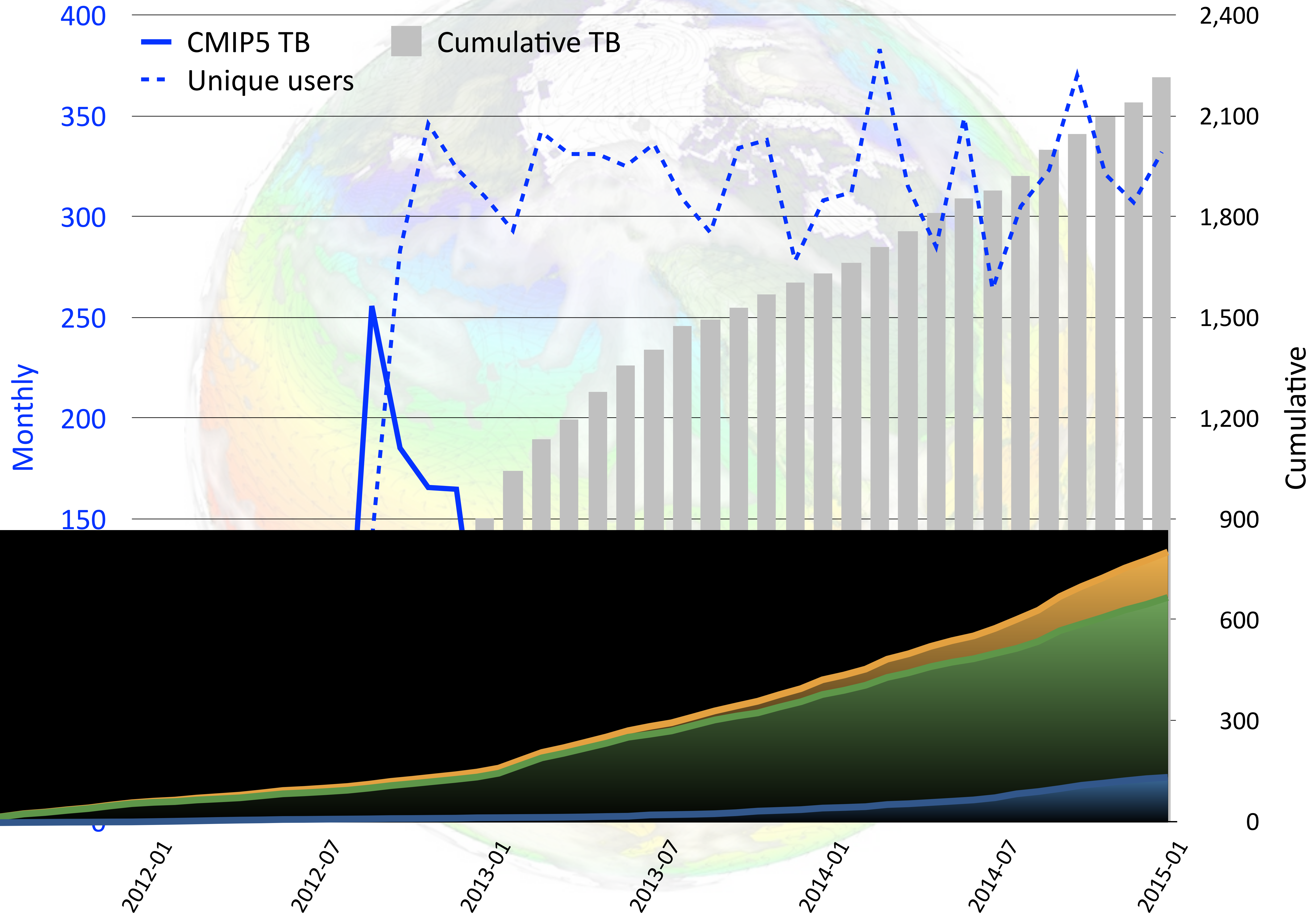


```
1999/09/08 392902052 /PCM1/pcm/IPCC/IPCC_BAU.2000-2099.nc
1999/09/09 155983196 /PCM1/pcm/IPCC/IPCC_Hist.1960-1999.nc
1999/09/16     66948 /PCM1/pcm/IPCC/gridinfo.nc
1999/11/19 530859560 /PCM1/pcm/IPCC/IPCC_Control_y000-149.nc
1999/11/19 530859676 /PCM1/pcm/IPCC/IPCC_Control_y150-299.nc
2000/02/01 372585720 /PCM1/pcm/IPCC/IPCC_STAB.2005-2099.nc
```

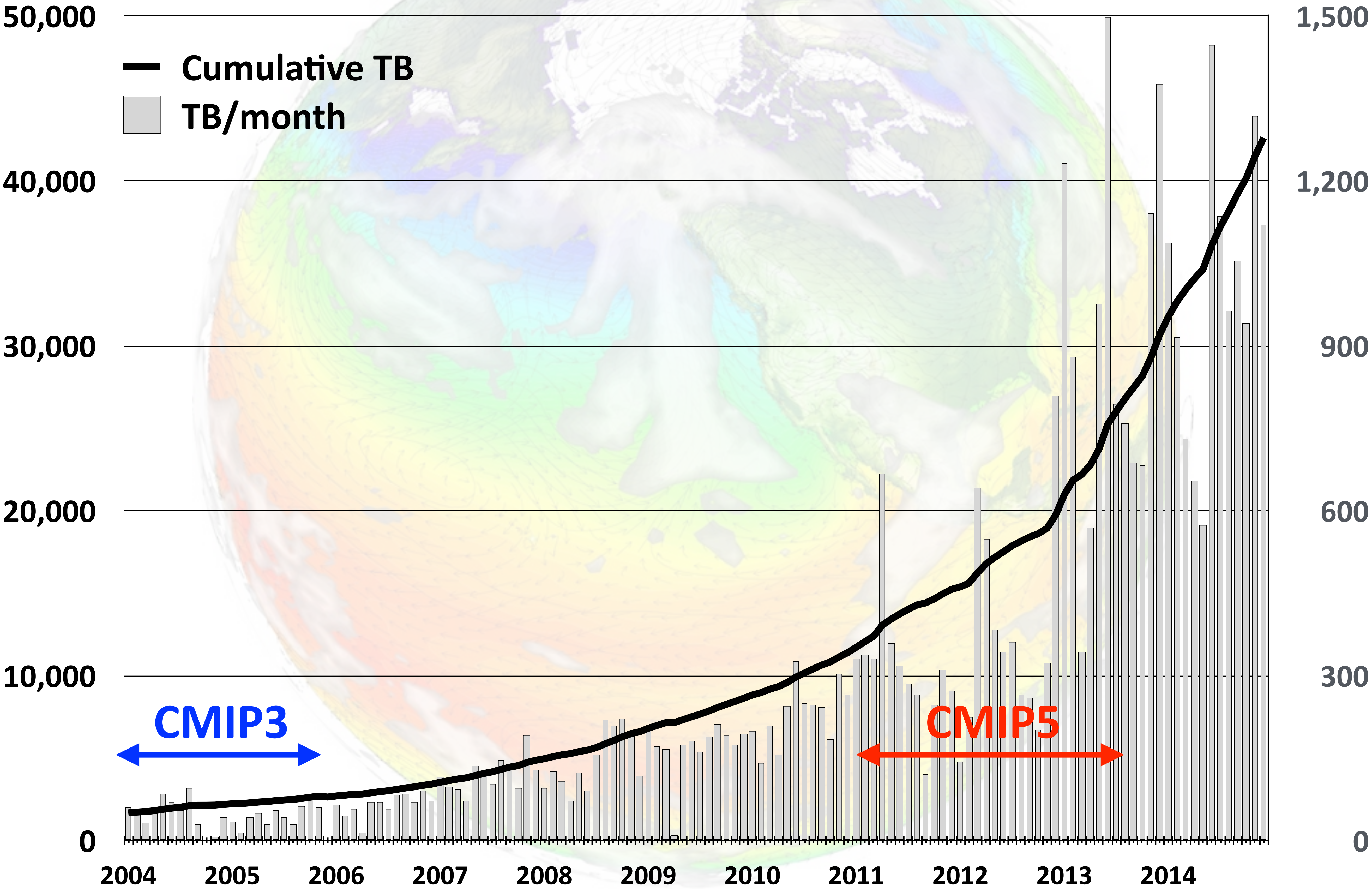
Metrics


900,000 GB
800,000 GB
700,000 GB
600,000 GB
500,000 GB
400,000 GB
300,000 GB
200,000 GB
100,000 GB
0 GB
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
TOTAL
CMIP3
NCAR

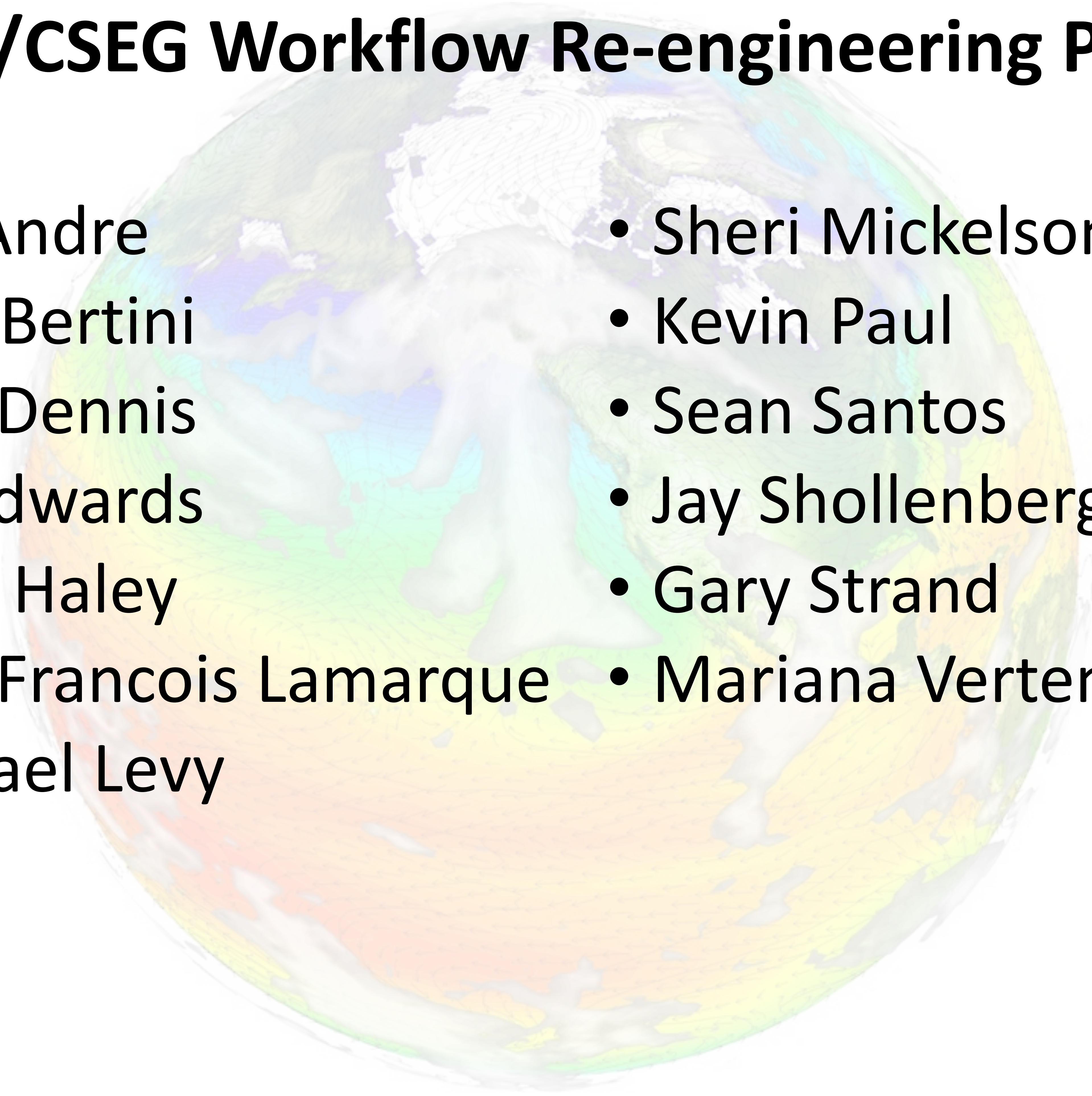
NCAR CMIP5 downloads metrics



NCAR archival storage



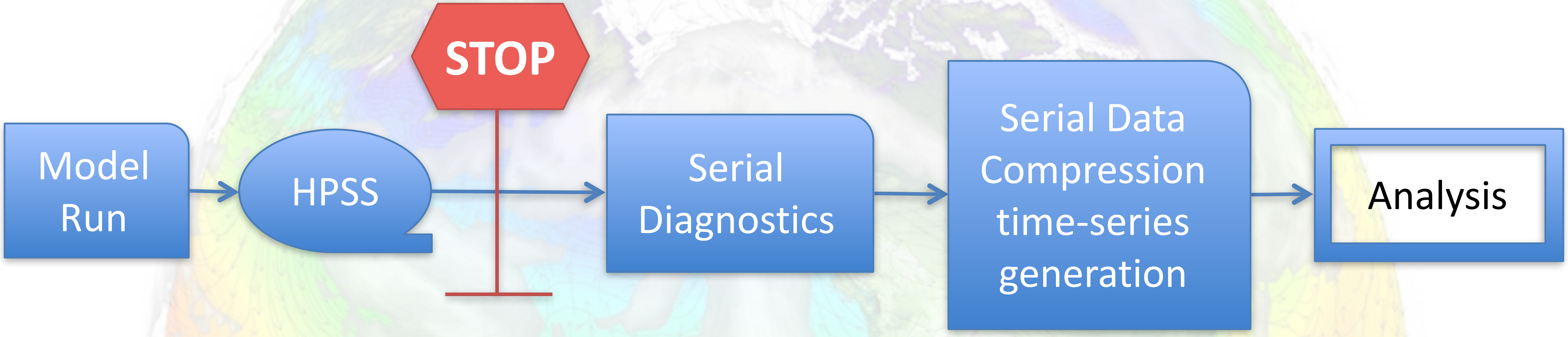
CESM/CSEG Workflow Re-engineering Project



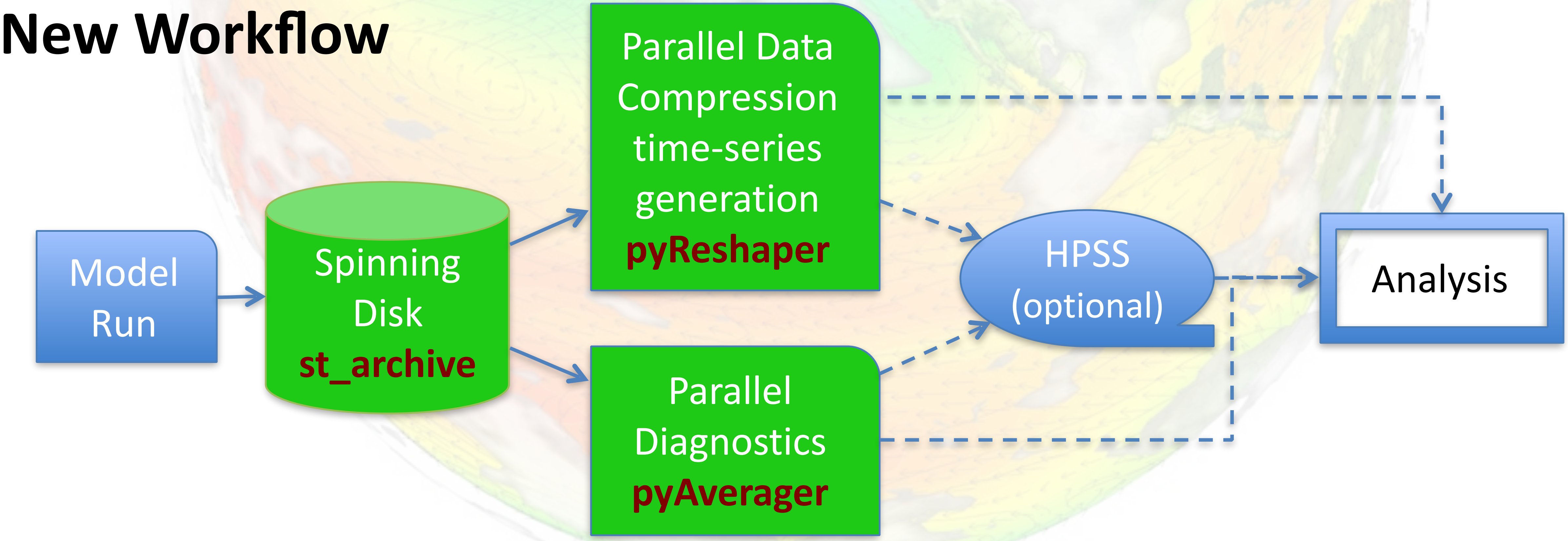
- Ben Andre
- Alice Bertini
- John Dennis
- Jim Edwards
- Mary Haley
- Jean-Francois Lamarque
- Michael Levy
- Sheri Mickelson
- Kevin Paul
- Sean Santos
- Jay Shollenberger
- Gary Strand
- Mariana Vertenstein

CESM/CSEG Workflow Re-engineering Project

Old Workflow

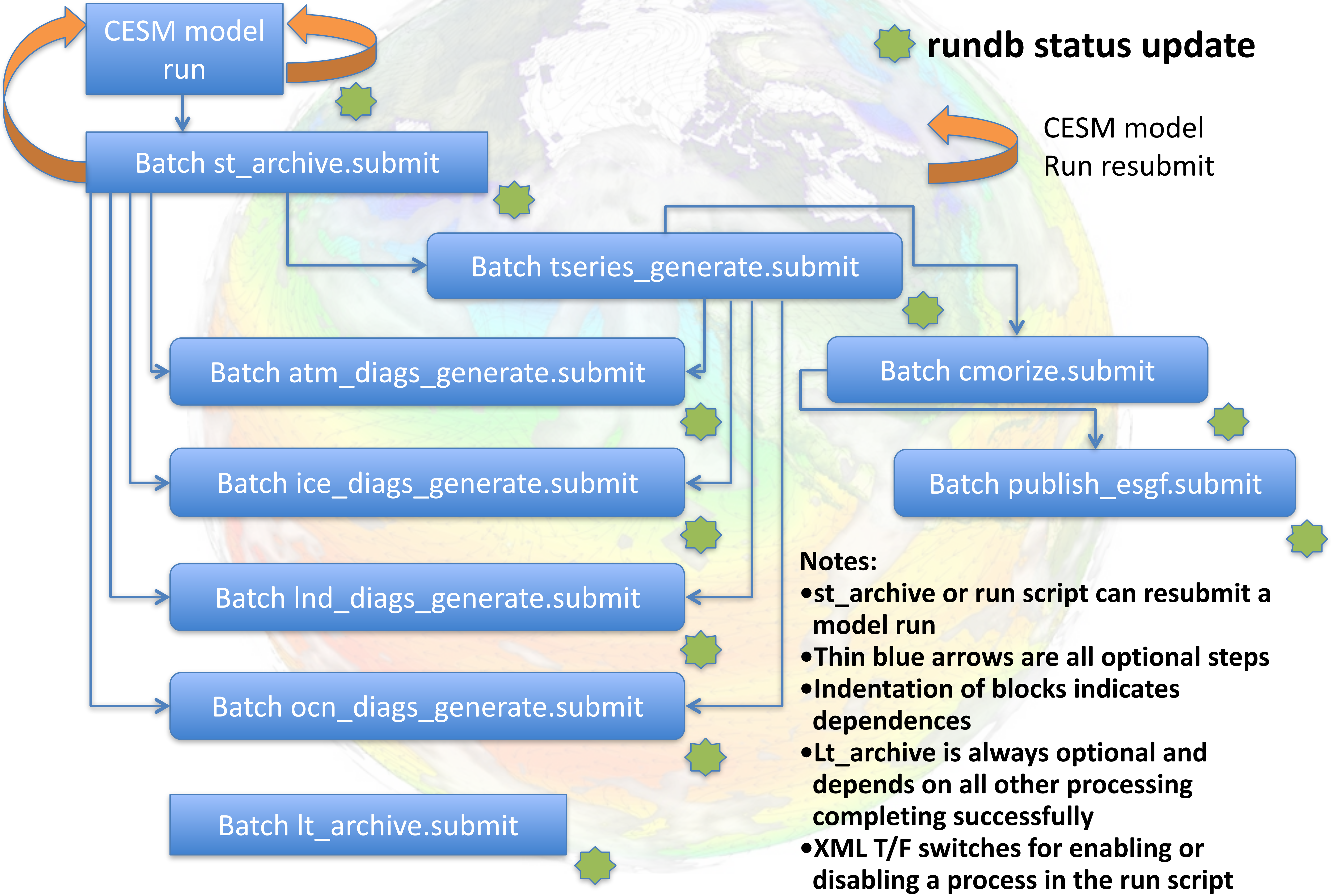


New Workflow



(Courtesy Sheri Mickelson, NCAR)

CESM/CSEG Workflow Re-engineering Project



rundb status update

CESM model
Run resubmit

CESM model run

Batch st_archive.submit

Batch tseries_generate.submit

Batch atm_diags_generate.submit

Batch ice_diags_generate.submit

Batch lnd_diags_generate.submit

Batch ocn_diags_generate.submit

Batch lt_archive.submit

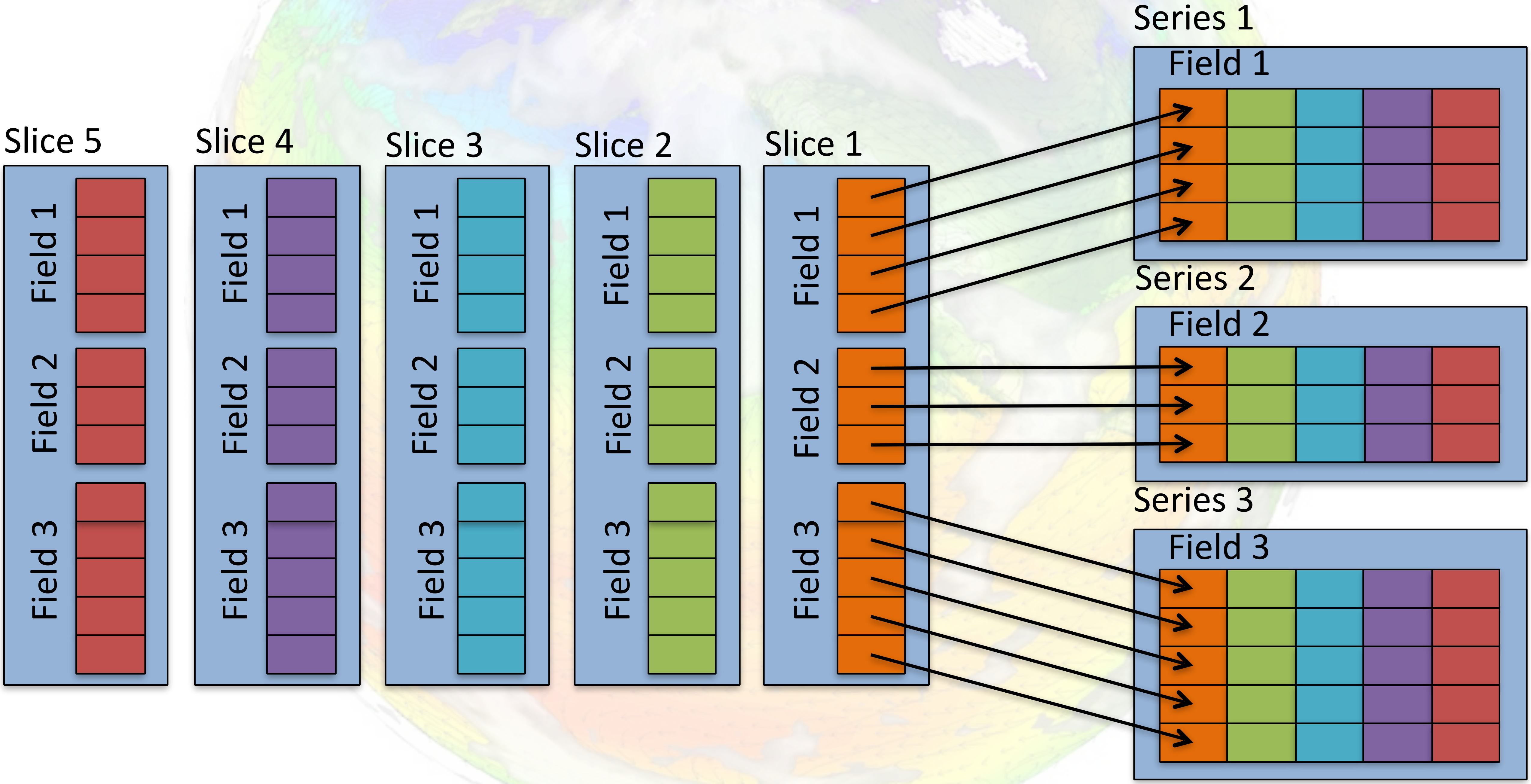
Batch cmorize.submit

Batch publish_esgf.submit

Notes:

- **st_archive or run script can resubmit a model run**
- **Thin blue arrows are all optional steps**
- **Indentation of blocks indicates dependences**
- **Lt_archive is always optional and depends on all other processing completing successfully**
- **XML T/F switches for enabling or disabling a process in the run script**

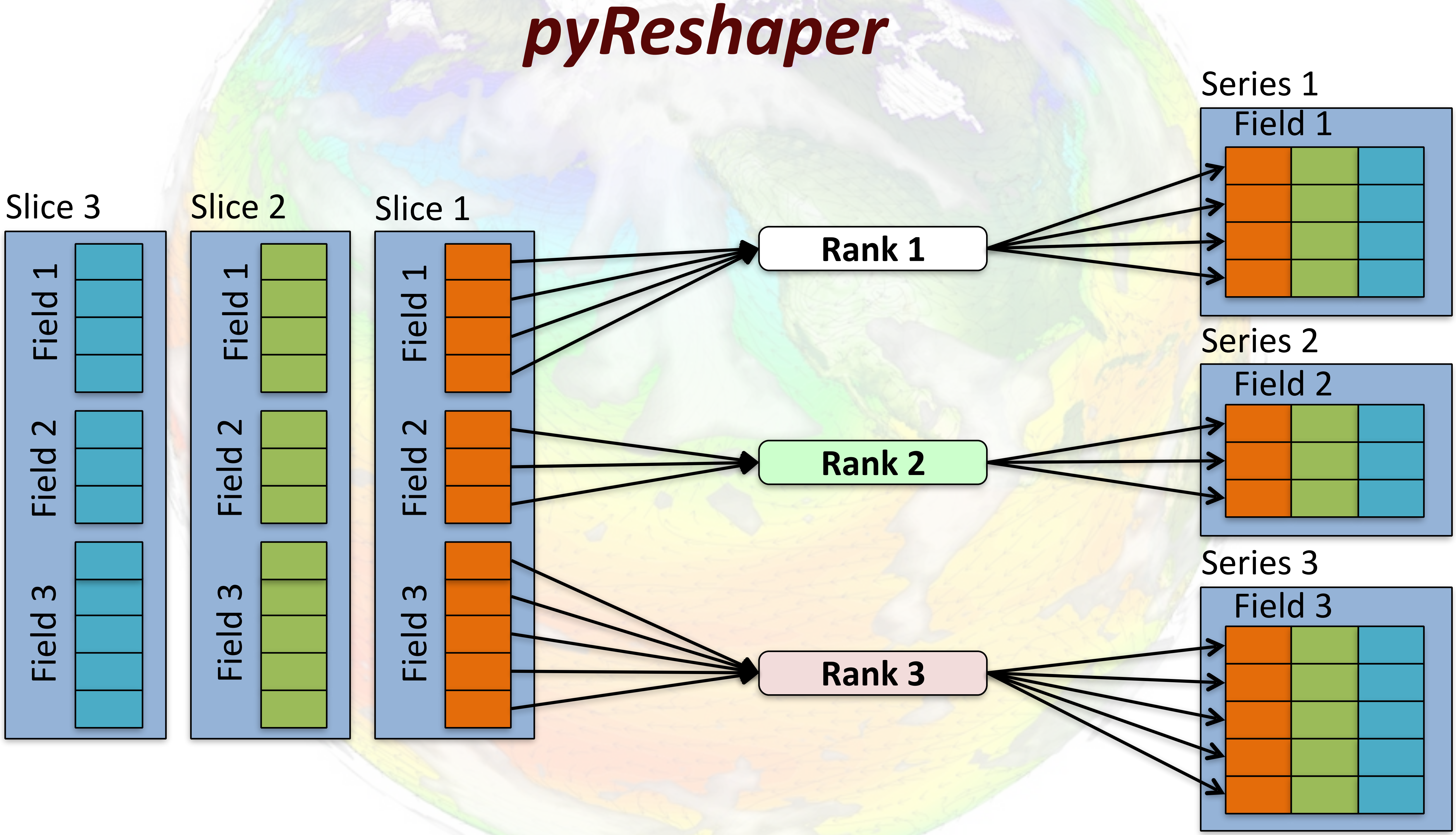
History Time-Slice to Time-Series Converter (Serial NCO)



(Courtesy Sheri Mickelson, NCAR)

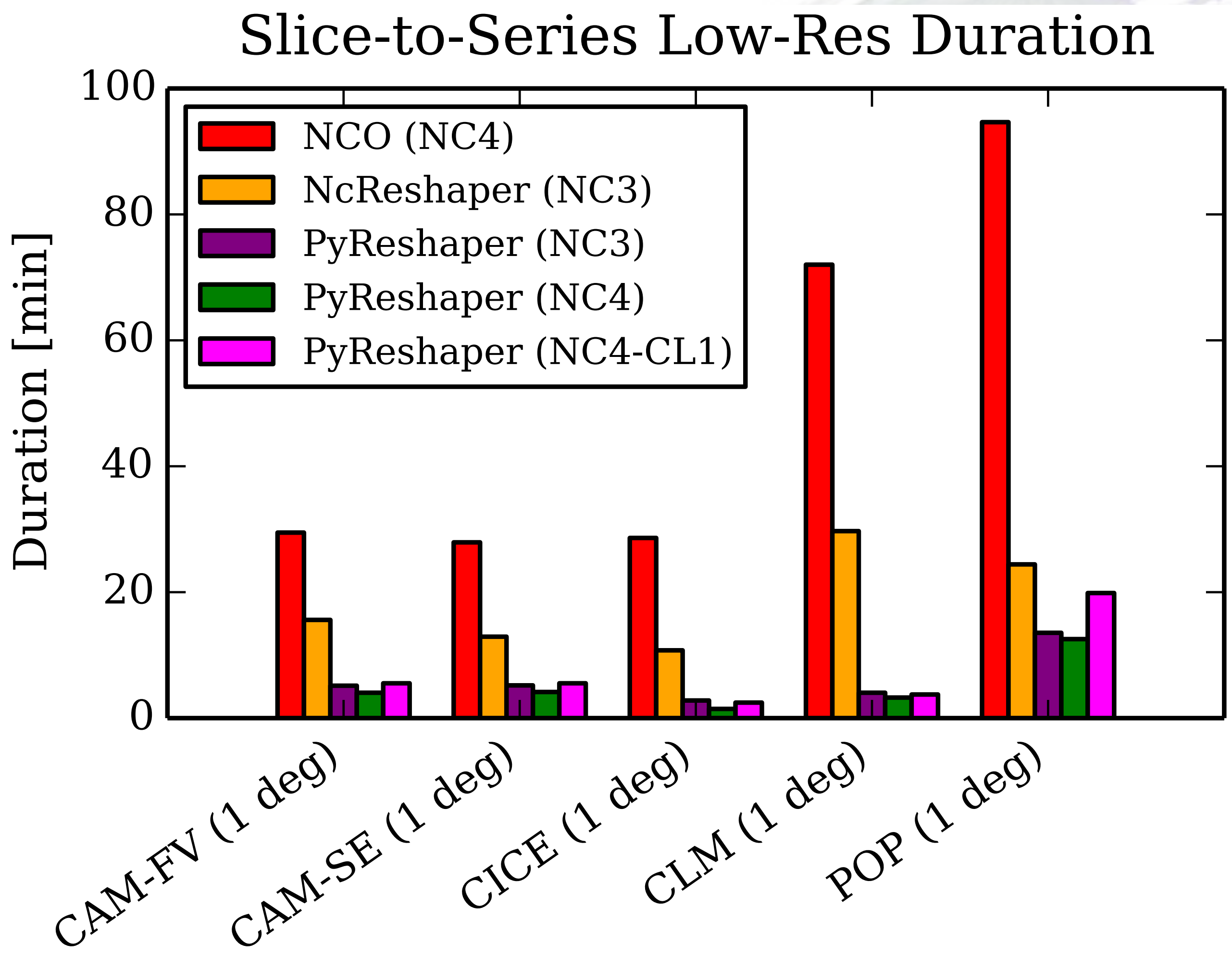
Task Parallelization Strategy

Each rank is responsible for writing one+ time-series variables to a file

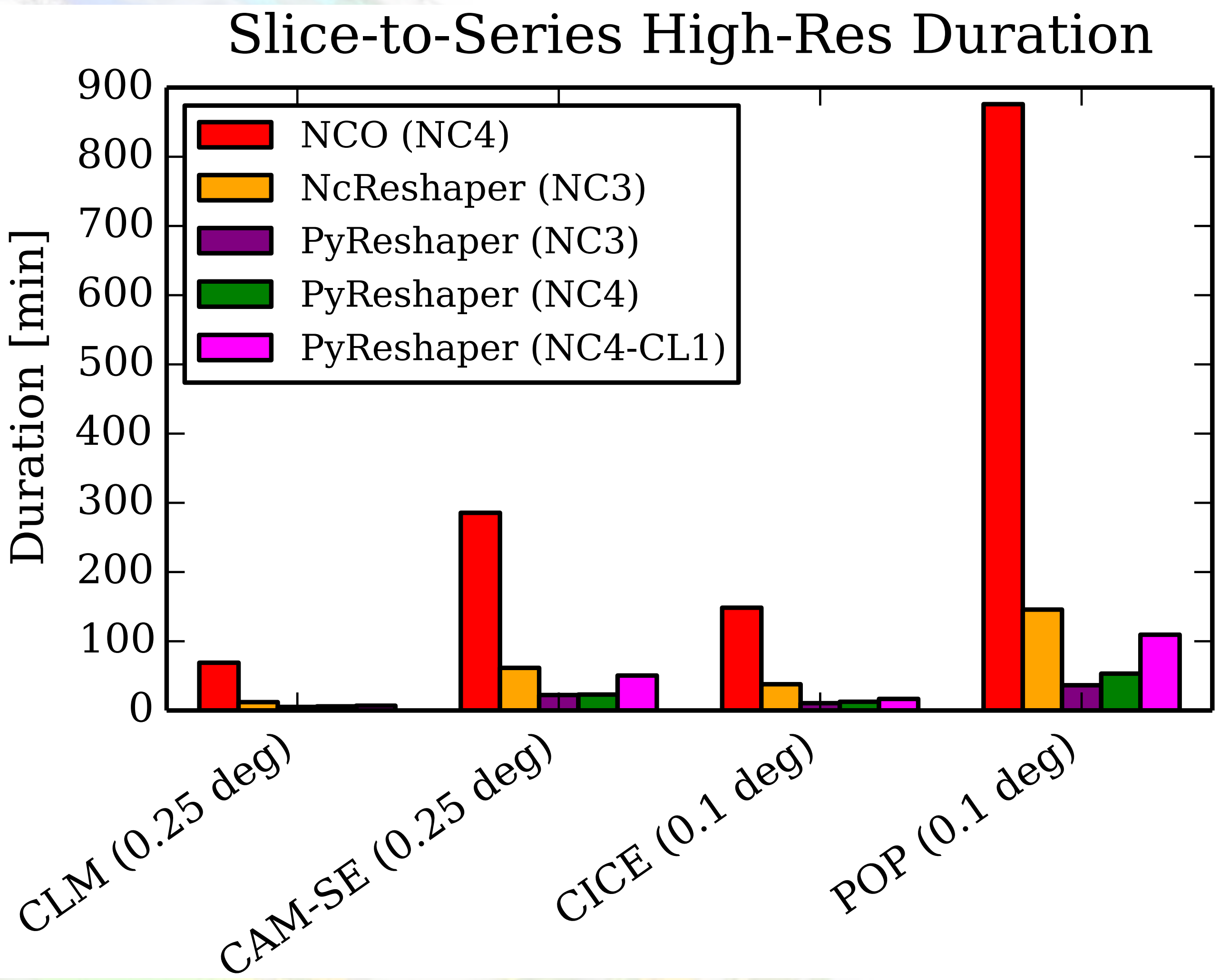


(Courtesy Sheri Mickelson, NCAR)

Time-Series Generation Performance



Slice-to-Series Low-Res Duration



Slice-to-Series High-Res Duration

Details from 1deg POP run:

- 10 years of monthly history data
- TI Metadata Variables: 63
- TV Metadata Variables: 2
- Time-Series Variables: 114
- Variables (TOTAL): 179

pyReshaper operated 4.5 times faster than NCO serial

Details from 0.1deg POP run:

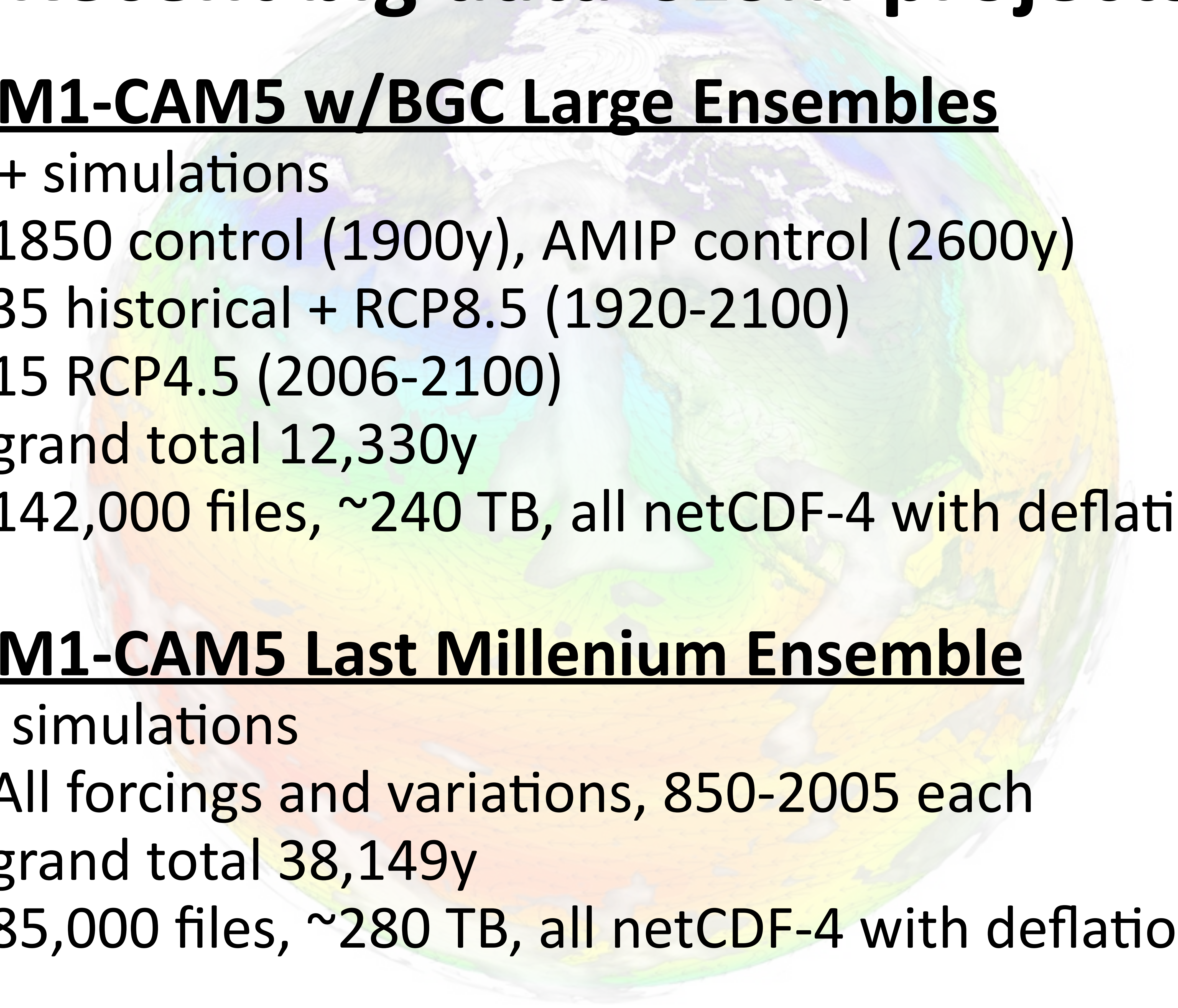
- 10 years of monthly history data
- TI Metadata Variables: 58
- TV Metadata Variables: 2
- Time-Series Variables: 34
- Variables (TOTAL): 94

pyReshaper operated 9 times faster than NCO serial

Yellowstone: 4 nodes & 4 cores/node.

(Courtesy Sheri Mickelson, NCAR)

Recent big data CESM projects



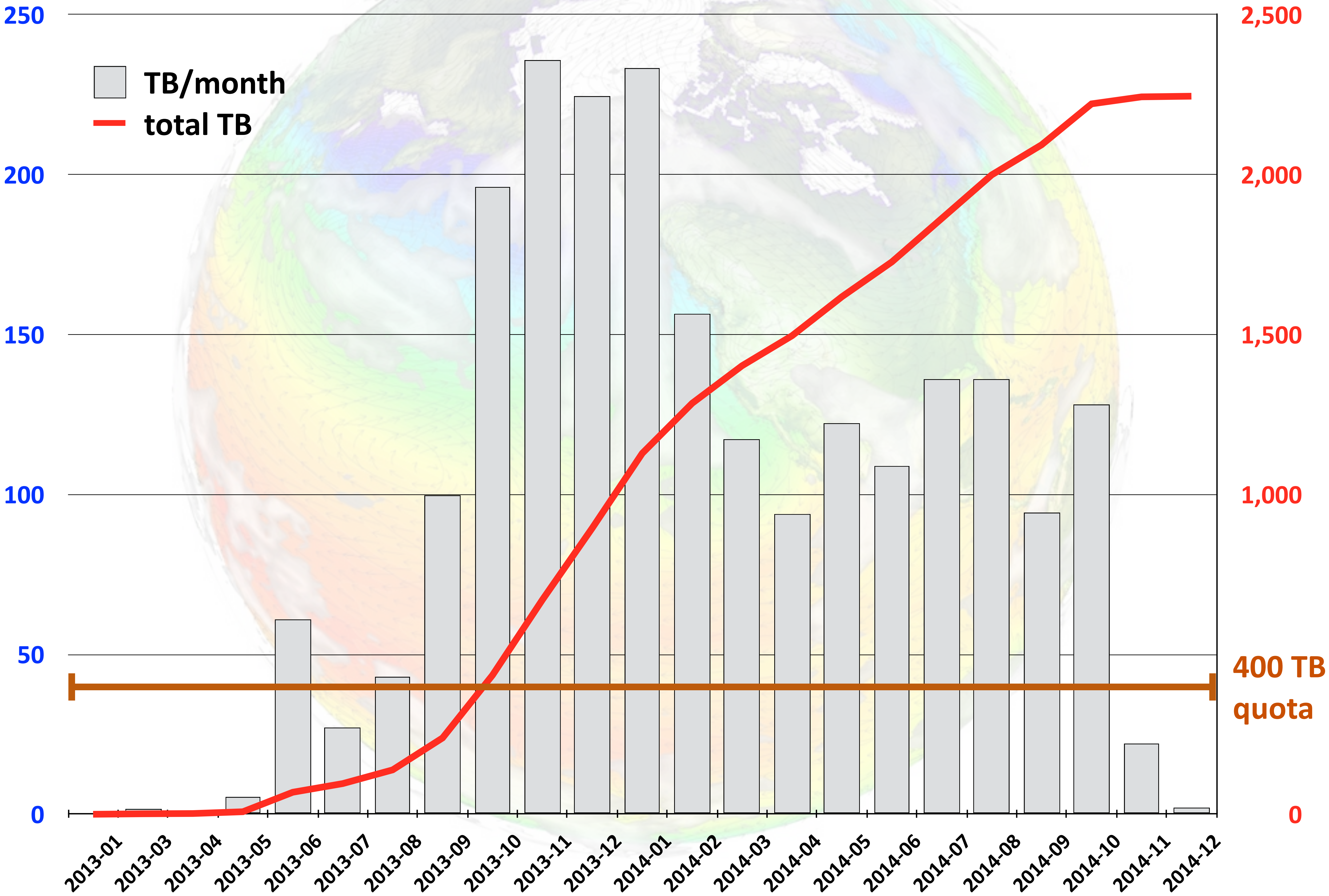
CESM1-CAM5 w/BGC Large Ensembles

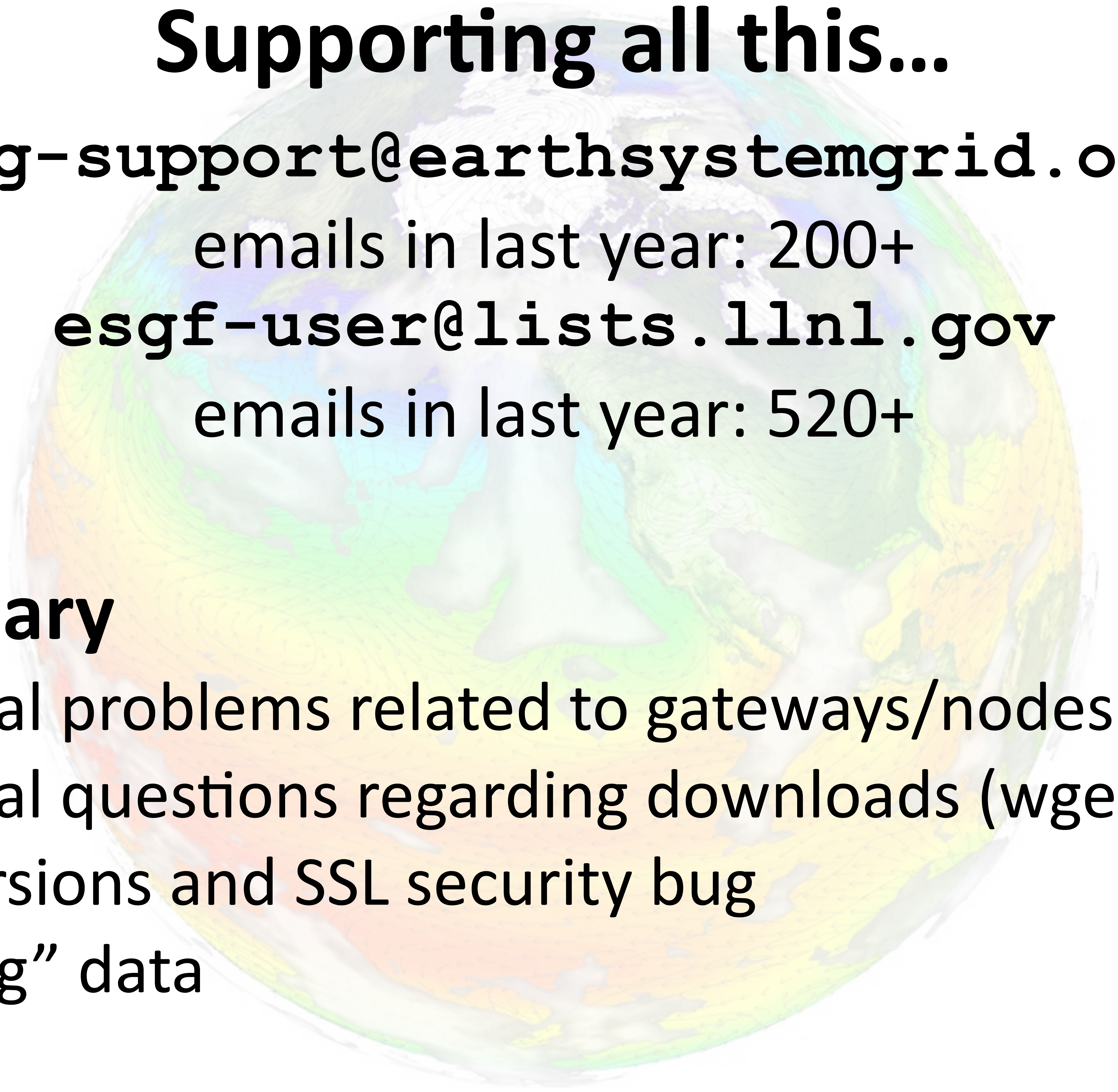
- 70+ simulations
  - 1850 control (1900y), AMIP control (2600y)
  - 35 historical + RCP8.5 (1920-2100)
  - 15 RCP4.5 (2006-2100)
  - grand total 12,330y
  - 142,000 files, ~240 TB, all netCDF-4 with deflation

CESM1-CAM5 Last Millenium Ensemble

- 33 simulations
  - All forcings and variations, 850-2005 each
  - grand total 38,149y
  - 85,000 files, ~280 TB, all netCDF-4 with deflation

Workflow-α test cases





Supporting all this…

esg-support@earthsystemgrid.org

emails in last year: 200+

esgf-user@lists.llnl.gov

emails in last year: 520+

Summary

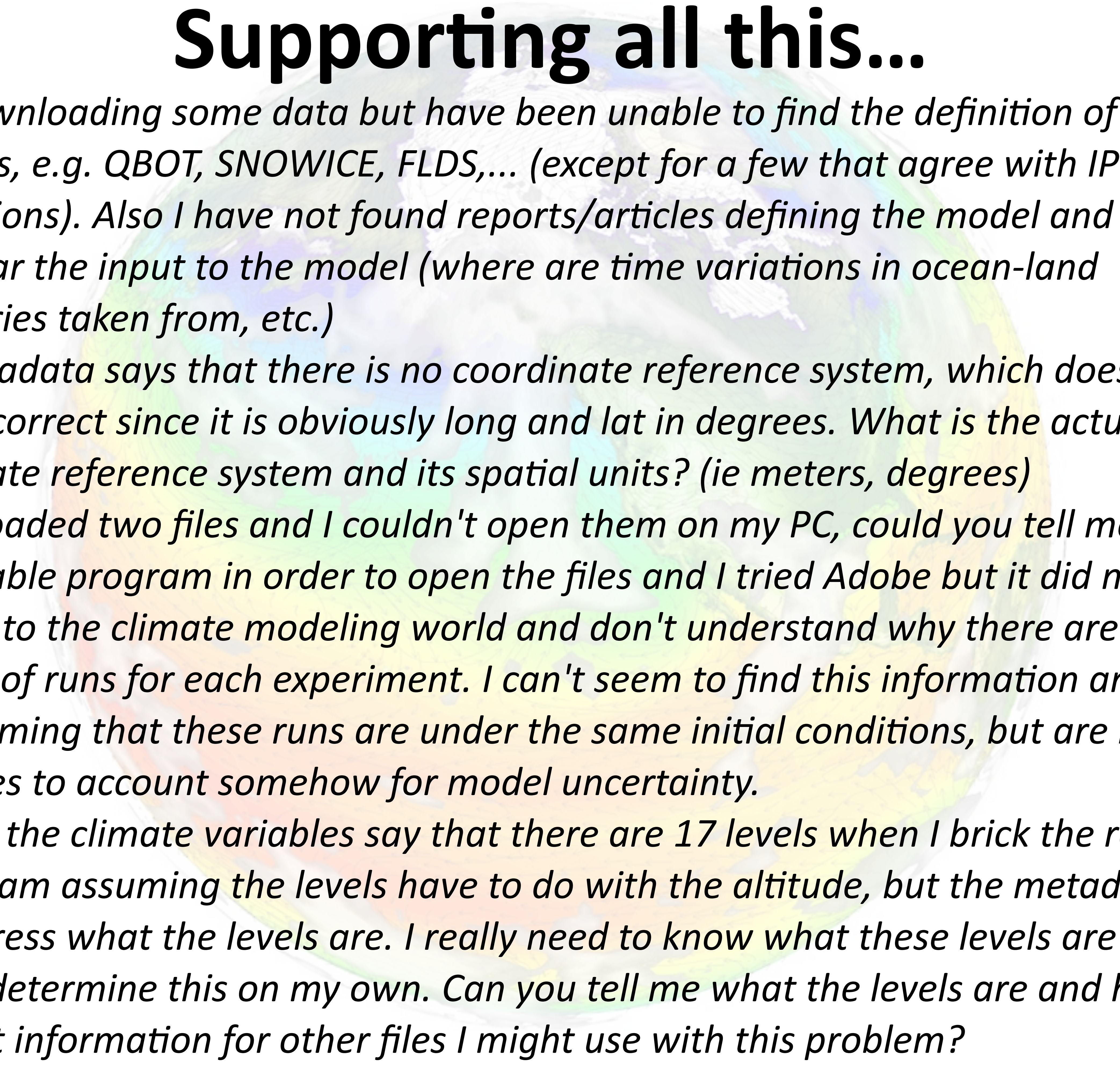
Technical problems related to gateways/nodes

Technical questions regarding downloads (wget, etc.)

Java versions and SSL security bug

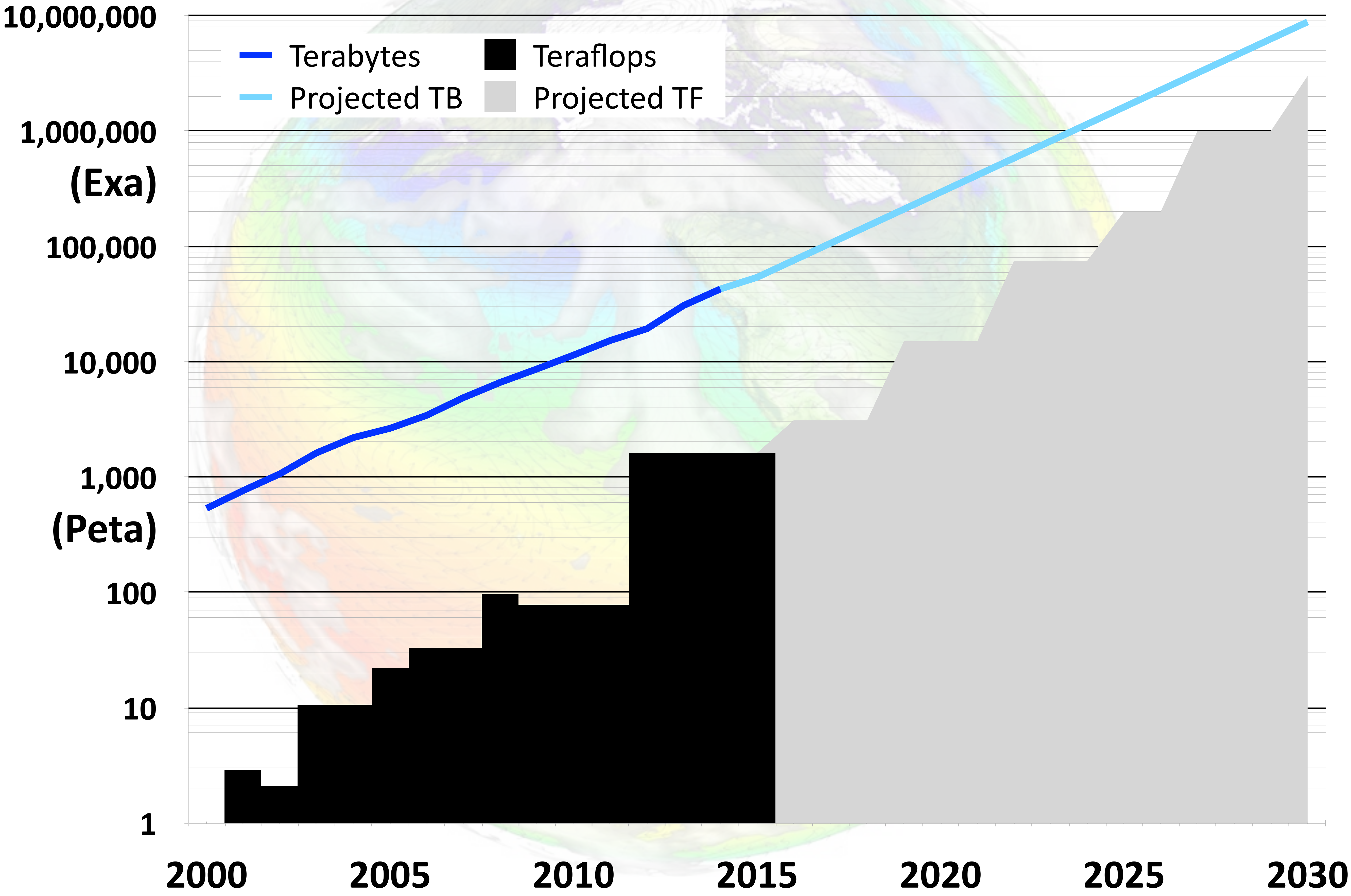
“Missing” data

Supporting all this…

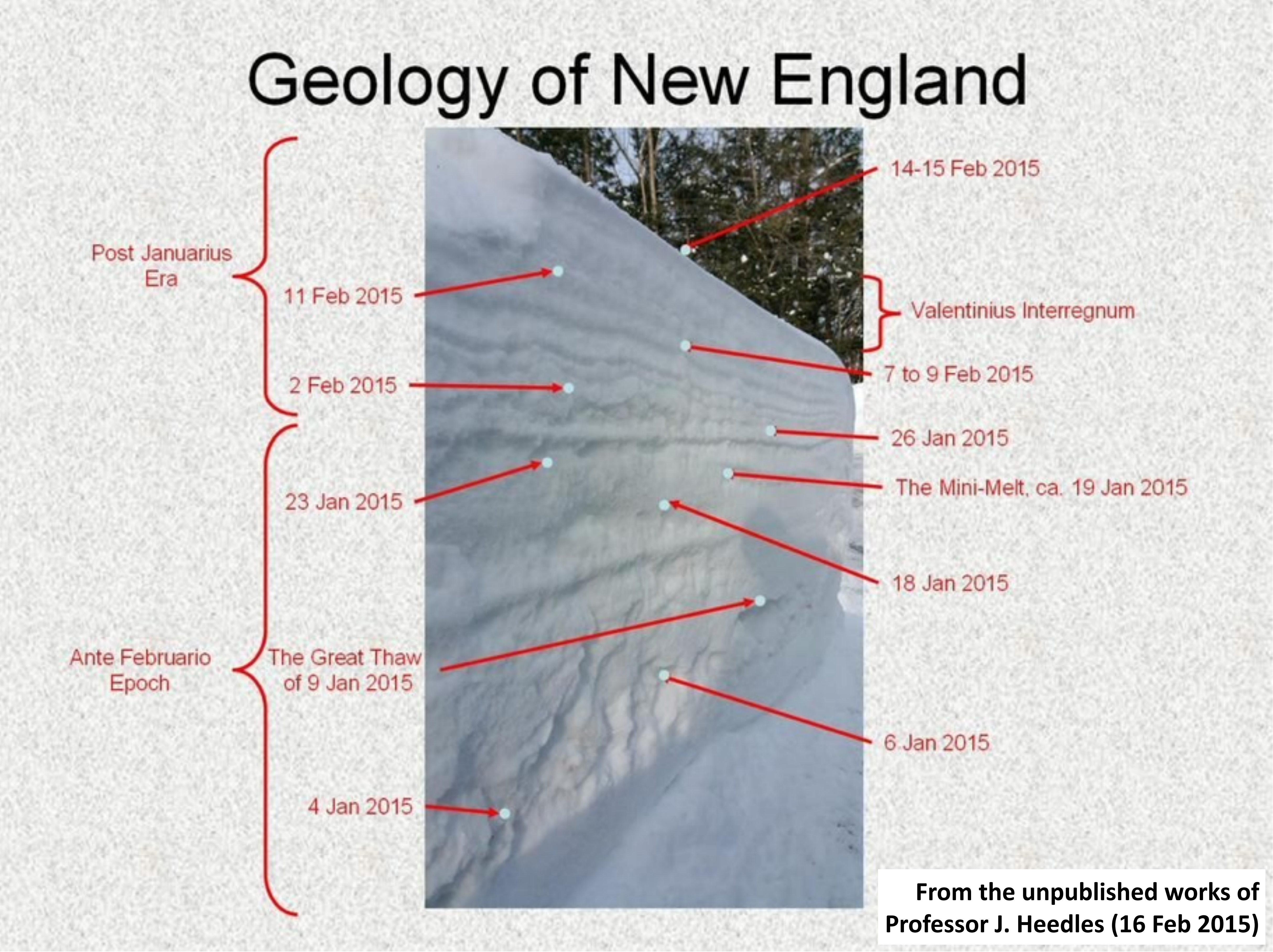


- *I am downloading some data but have been unable to find the definition of variables, e.g. QBOT, SNOWICE, FLDS,... (except for a few that agree with IPCC conventions). Also I have not found reports/articles defining the model and in particular the input to the model (where are time variations in ocean-land boundaries taken from, etc.)*
- *The metadata says that there is no coordinate reference system, which doesn't appear correct since it is obviously long and lat in degrees. What is the actual coordinate reference system and its spatial units? (ie meters, degrees)*
- *I downloaded two files and I couldn't open them on my PC, could you tell me what is the suitable program in order to open the files and I tried Adobe but it did not work.*
- *I'm new to the climate modeling world and don't understand why there are a number of runs for each experiment. I can't seem to find this information anywhere. I'm assuming that these runs are under the same initial conditions, but are maybe replicates to account somehow for model uncertainty.*
- *Some of the climate variables say that there are 17 levels when I brick the raster into R. I am assuming the levels have to do with the altitude, but the metadata does not address what the levels are. I really need to know what these levels are and how to determine this on my own. Can you tell me what the levels are and how to find that information for other files I might use with this problem?*

NCAR flops and bytes 2000-2030



Just for fun



Geology of New England

Post Januarius Era

11 Feb 2015

2 Feb 2015

Ante Februario Epoch

23 Jan 2015

The Great Thaw of 9 Jan 2015

4 Jan 2015

14-15 Feb 2015

Valentinius Interregnum

7 to 9 Feb 2015

26 Jan 2015

The Mini-Melt, ca. 19 Jan 2015

18 Jan 2015

6 Jan 2015

From the unpublished works of
Professor J. Heedles (16 Feb 2015)